## 2015 Joint UNECE/Eurostat Work Session on SDC

# Transparency and microaggregation

Vicenç Torra[1]

October, 2015

[1] University of Skövde, Sweden

# Outline

# Transparency

## Transparency

# Transparency

## Transparency.

- "the release of information about processes and even parameters used to alter data" (Karr, 2009).

## Effect.

- Information Loss. Positive effect, less loss
  E.g., noise addition $\rho(X) = X + \epsilon$ where $\epsilon$ s.t.
  $E(\epsilon) = 0$ and $Var(\epsilon) = kVar(X)$

$$Var(X') = Var(X) + kVar(X) = (1 + k)Var(X).$$

# Transparency

**Transparency.**

- "the release of information about processes and even parameters used to alter data" (Karr, 2009).

**Effect.**

- Disclosure Risk. Negative effect, larger risk
  - Attack to single-ranking microaggregation (Winkler, 2002)
  - Formalization of the transparency attack (Nin, Herranz, Torra, 2008)
  - Attacks to microaggregation and rank swapping (Nin, Herranz, Torra, 2008)

# Transparency

## Transparency.

- "the release of information about processes and even parameters used to alter data" (Karr, 2009).

## Effect.

- Disclosure Risk. Formalization (Nin, Herranz, Torra, 2008)
  - $X$ and $X'$ original and masked files, $\mathbf{V} = (V_1, \ldots, V_s)$ attributes
  - $B_j(x)$ set of masked records associated to $x$ w.r.t. $j$th variable.
  - Then, for record $x$, the masked record $x_\ell$ corresponding to $x$ is in the intersection of $B_j(x)$.

$$x_\ell \in \cap_j B_j(x).$$

- Worst case scenario in record linkage: upper bound of risk

# Transparency

**Transparency.**

- "the release of information about processes and even parameters used to alter data" (Karr, 2009).

**Need of methods resistant to transparency attacks.**

- p-buckets and p-distribution rank swapping (Nin, Herranz, Torra, 2008)
- fuzzy microaggregation

# Fuzzy Microaggregation

## Fuzzy microaggregation

# Fuzzy microaggregation

**Approach.** Use fuzzy clustering: fuzzy $c$-means

- Cluster the data set using $c$ and $m$
  - $c$: number of clusters
  - $m$: fuzzy degree
- Procedure: iterative method
  - Compute cluster centers
  - Compute membership (assignment of objects to clusters)

# Fuzzy microaggregation

**Approach.** Use fuzzy clustering: fuzzy $c$-means

- Cluster the data set using $c$ and $m$
  - $c$: number of clusters
  - $m$: fuzzy degree
    - ⋆ The larger the $m$,
      the larger the overlapping among cluster centers.
    - ⋆ The larger the $m$,
      the nearer the cluster centers to the center of the data

# Fuzzy microaggregation

**Approach.** Use fuzzy clustering: fuzzy $c$-means

- Cluster the data set using $c$ and $m$
  - $c$: number of clusters

$$c \in \left[\frac{|X|}{2k}, \frac{|X|}{k}\right].$$

Therefore, average number of records per cluster

$$k < |X|/c < 2k.$$

  - $m$: fuzzy degree: $\rightarrow$ use two $m_1$ and $m_2$

# Fuzzy microaggregation

**Approach.** Use fuzzy clustering: fuzzy $c$-means

- Cluster the data set using $c$ and $m$
  - $c$: number of clusters

$$c \in \left[\frac{|X|}{2k}, \frac{|X|}{k}\right].$$

Therefore, average number of records per cluster

$$k < |X|/c < 2k.$$

  - $m$: fuzzy degree: $\rightarrow$ use two $m_1$ and $m_2$
- Procedure: iterative method
  - Compute cluster centers: use $m_1$
  - Compute membership: use $m_2$

# Fuzzy microaggregation

**Approach.** Use fuzzy clustering: fuzzy $c$-means

- Cluster the data set using $c$ and $m$
  - $c$: number of clusters

$$c \in \left[\frac{|X|}{2k}, \frac{|X|}{k}\right].$$

Therefore, average number of records per cluster

$$k < |X|/c < 2k.$$

  - $m$: fuzzy degree: $\rightarrow$ use two $m_1$ and $m_2$

# Fuzzy microaggregation

**Approach.** Use fuzzy clustering: fuzzy $c$-means

- Cluster the data set using $c$ and $m$
  - $c$: number of clusters

$$c \in \left[\frac{|X|}{2k}, \frac{|X|}{k}\right].$$

Therefore, average number of records per cluster

$$k < |X|/c < 2k.$$

  - $m$: fuzzy degree: $\rightarrow$ use two $m_1$ and $m_2$
- Procedure: iterative method
  - Compute cluster centers.

    use $m_1$: the larger $m_1$, nearer cluster centers to center of the data
  - Compute membership.

    use $m_2$: the larger $m_2$, the larger the overlapping

# Fuzzy microaggregation

**Step 1:** Apply fuzzy $c$-means with a given $c$ and a given $m := m_1$.

**Step 2:** Compute memberships $u$ for $x_k$ and all clusters $i$ given $m_2$.

$$u_{ik} = \left( \sum_{j=1}^{c} \left( \frac{||x_k - v_i||^2}{||x_k - v_j||^2} \right)^{\frac{1}{m_2 - 1}} \right)^{-1}$$

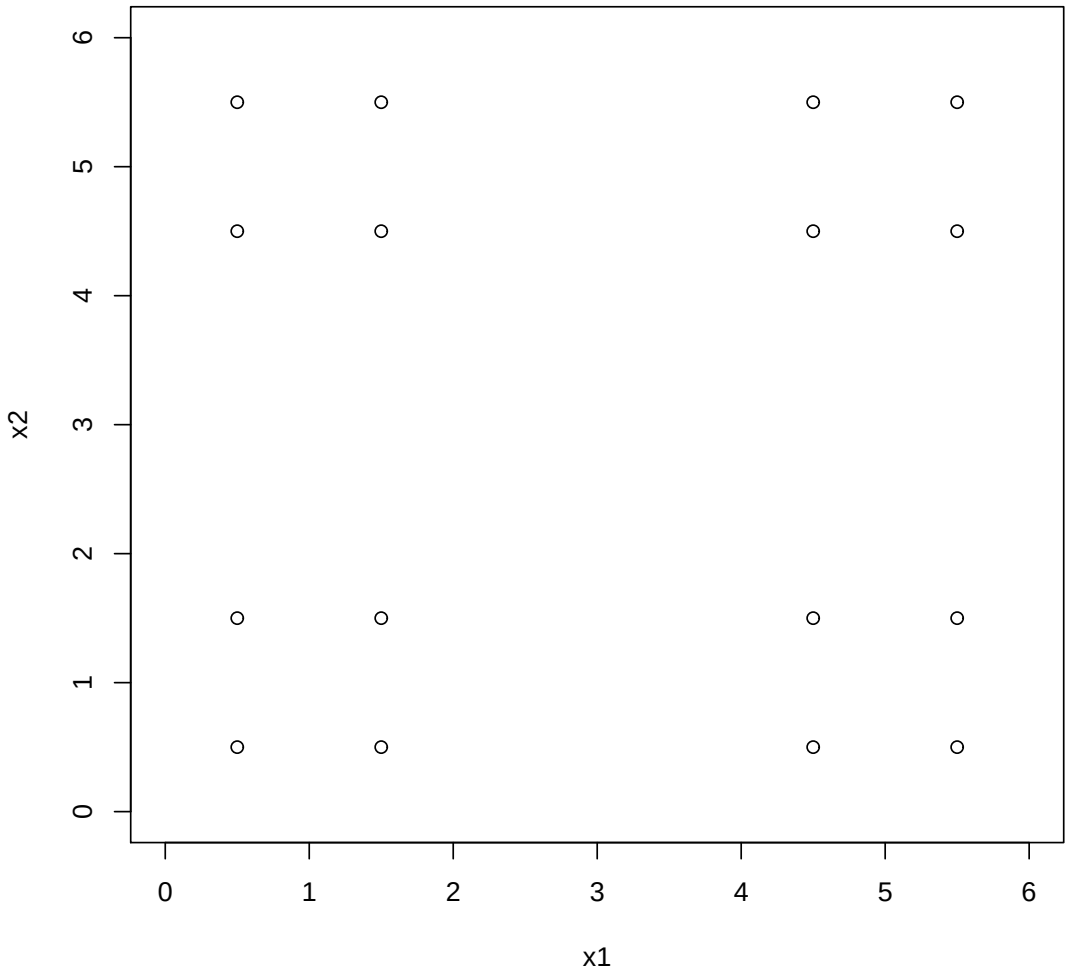
**Step 3:** Assign each $x_k$ to a cluster

determine a random value $\chi$ in $[0, 1]$,

assign $x_k$ to the $i$th cluster using probability distribution $u_{1k}, \ldots, u_{ck}$

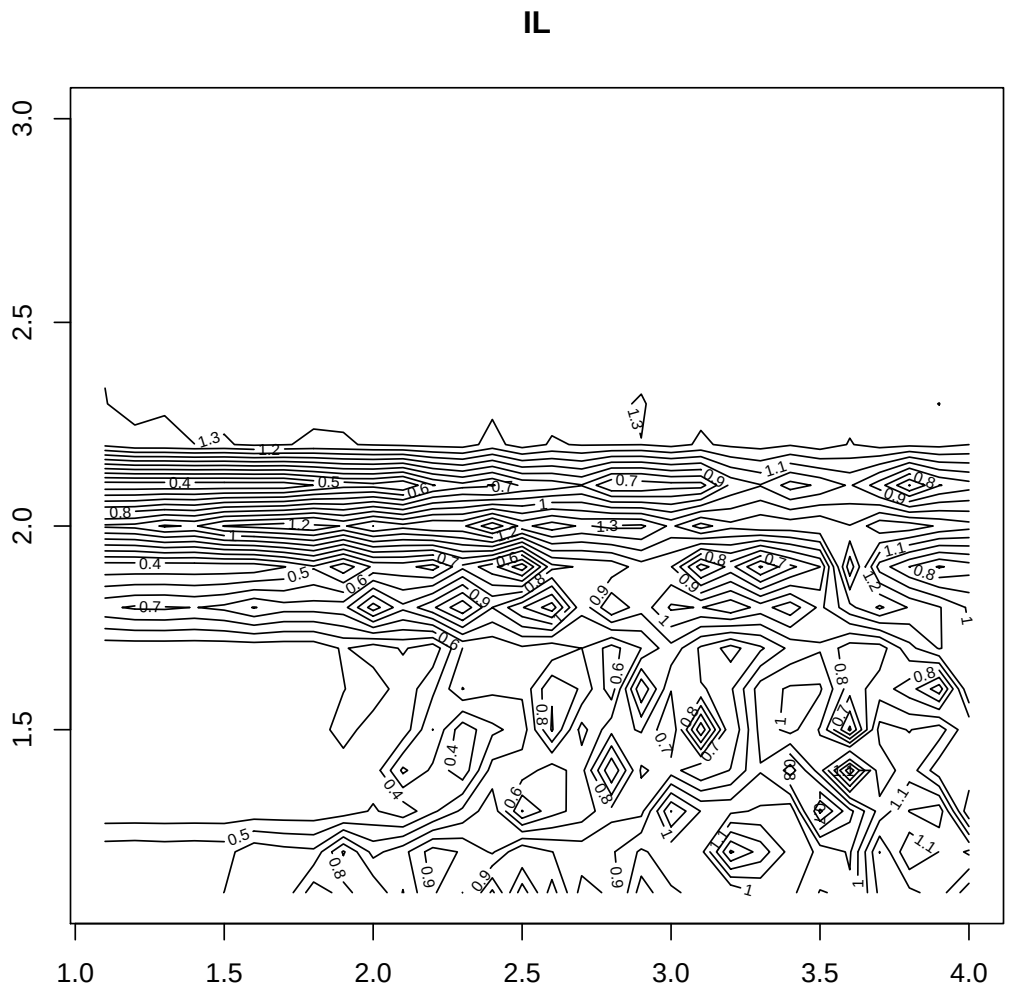
# Fuzzy microaggregation

**Example.** 16 points



# Fuzzy microaggregation

**Example.** 16 points



# Fuzzy microaggregation

**Analysis.**

1. The larger the $m_1$, the larger the information loss.
2. The larger the $m_2$, the larger the information loss.
3. The smaller the number of clusters $c$, the larger the information loss.
4. It does not ensure $k$ indistinguishable records, but
   for appropriate $m_2$, no one-to-one correspondence between the record and the cluster center used for its replacement.
   for large $m_2$ assignment is equally probable to all clusters.
   $\rightarrow$ on average the outcome satisfies $k$-anonymity

# Summary

## Summary

# Summary

## Conclusions.

- Transparency improves data quality,
- Transparency increases risk.
- Masking methods can be designed to avoid transparency attacks
- Fuzzy microaggregation to avoid transparency attacks

# Thank you